A simplified approach to generating synthetic data for disclosure control

Gillian M. Raab*, Beata Nowok† and Chris Dibben‡

School of Geosciences, University of Edinburgh, Edinburgh, UK

March 10, 2015

Abstract

We describe results on the creation and use of synthetic data that were derived in the context of a project to make synthetic extracts available for users of the UK Longitudinal Studies. Contrary to the existing literature we show that there are circumstances when inferences can be made from fully synthetic data generated from fitted parameters without sampling from their posterior distributions (simple synthesis). The condition that allows this, which we describe as "common-sampling", is that the original sample and the synthetic data can be considered as sampled in the same way from their respective populations. New variance estimators for the analysis of synthetic data are derived when the common-sampling condition is met. It is shown that simple synthesis, with these estimators, provide better estimates than the methods suggested in the literature for fully synthetic data. The results are confirmed by simulations and are illustrated with an example from the Scottish Longitudinal Study.

Keywords— synthetic data, disclosure control, CART, UK Longitudinal Studies

*gillian.raab@ed.ac.uk
†beata.nowokb@ed.ac.uk
‡chris.dibben@ed.ac.uk

1 Introduction and background

1.1 Synthetic data for disclosure control

National statistics agencies and other groups collect large amounts of information about individuals and organisations. Such data can be used to understand population processes so as to inform policy and planning. The cost of such data can be considerable, both for the collectors and the subjects who provide their data. Because of confidentiality constraints and guarantees to data subjects full access to such data is often restricted to the staff of the collection agencies. Traditionally, data collectors have used anonymization along with simple perturbation methods such as aggregation, top-coding, record-swapping, suppression of sensitive cells or adding random noise, to prevent the identification of data subjects. Advances in computer technology and search techniques have shown that such measures may not prevent disclosure (Ohm (2010)). Additionally, examples have shown that these ad hoc disclosure procedures may compromise the conclusions that can be drawn from such data (Winkler (2007), Elliot and Purdam (2007)).

In response to these difficulties there have been several initiatives e.g. US Census Bureau (2011) and US Census Bureau (2013) to provide synthetic data to facilitate researchers' access to confidential data. Synthetic data reproduce the essential features of the actual data with some or all of the values replaced by simulations from probability distributions. The monograph by Drechsler (2011b) summarises some of the theoretical and policy developments. Methods are detailed in Raghunathan et al. (2003) and have been further discussed, extended and exemplified in a series of papers (Raghunathan et al. (2003); Reiter (2005a); Caiola and Reiter (2010); Dreschler and Reiter (2010); Reiter (2005b); Kinney and Reiter (2010); Dreschler (2012), among others). The aim of the methods, as described in these papers, has been to provide multiple synthetic data sets that can be used for inference. We envisage a different role for synthetic data which recognises its limitations. Staff with access to the original confidential data (synthesisers) create synthetic data which is made available to analysts. The analysts carry out data preparation, exploratory analyses and preliminary modelling on the synthetic data. The code developed on the synthetic data is used to run final fits of candidate models and validation analyses on the original un-synthesised data. The US Census Bureau initiatives, cited above, use this approach and the term "gold-standard" describes the final steps where code for fitting and validating models is run on the actual data.

The term "fully synthetic data" is used when all of the data are replaced by synthesised values, except perhaps for some variables, usually those defining the design, which are left unchanged. The results we present for inference from fully synthetic data must be conditional on the values of the unchanged data. Partially synthetic data are generated when only some of the data are synthesised. This may be either only certain selected cases or certain selected variables or perhaps selected cases from selected variables.

1.2 Synthesis compared to imputation

Data synthesis has been considered as an example of multiple imputation (Reiter and Raghunathan (2007)) and synthetic data is sometimes called "multiply imputed microdata", e.g. by Reiter (2005a). Although there are clear parallels between multiple imputation for missing data and data synthesis there are also important differences. Sequential versions of multiple imputation, first proposed by Rubin (1987), p.192 and developed by Raghunathan et al. (2001), van Buuren (2007) and others, simulate the missing values for each variable from a fit to the observed data, conditional on all the rest of the data, including the imputed values of other variables. To impute missing values for $y_1, y_2, ..., y_p$ models are fitted to the conditional distributions with parameters $\phi_1, \phi_2, ..., \phi_p$ denoted by $g_1(y_1|\phi_1, y_2, y_3, ..., y_p)$, $g_2(y_2|\phi_2, y_1, y_3, ..., y_p)$..., $g_p(y_p|\phi_p, y_1, y_2, ..., y_{p-1})$. Several iterations are required for the results to stabilise, after which the missing values are replaced by samples from their predictive distributions given the observed and imputed data. If the conditional distributions are not compatible (Arnold et al. (2001)) their joint distribution may not exist although the procedures have been shown to work well even with incompatible conditionals (van Buuren (2007)). For synthesis, the joint distribution used to generate the data is built up of distributions with parameters $\theta_1, \theta_2, ...\theta_p$ by starting with synthesis from one marginal distribution and then building up further distributions by conditioning on the variables already synthesised, e.g. $f_1(y_1|\theta_1)$, $f_2(y_2|\theta_2, y_1))$, $f_3(y_3|\theta_3, y_1, y_2))$,...,$f_3(y_p|\theta_p, y_1, ..., y_{p-1}))$. The product of these defines the joint distribution with parameter vector θ given by the concatenation of $\theta_1, \theta_2, ..., \theta_p$ and, except in some cases of partial synthesis, no iterations will be required.

Variance estimates calculated from data which include imputed values require an additional term to account for the loss of information from the missing data. To account for all of the uncertainty due to the missing data imputations are generating from posterior distributions which include the uncertainty in the fitted parameters of the models. The term "proper imputation" is used for this. For synthesis the situation is different; data are not truly missing and they inform the model used to generate synthetic data. The formulae for obtaining variances from synthetic data are therefore also quite different from those for imputation. We will show below that in some circumstances fully synthetic data can be generated from fits to the observed data and sampling from the posterior distributions of the parameters is not necessary.

For both multiple imputation and multiple synthesis, inferences will only be assured of validity if the model used for synthesis or imputation is the true one that generated the real data. Despite the theoretical inadequacies of sequential imputation it has been shown to be a useful practical tool (van Buuren (2007)) and to be relatively insensitive to model assumptions. This robustness of multiple imputation may well be due to the fact that only a small proportion of the data are replaced, a condition that will not hold for most synthetic data applications. Thus we expect correct model specification to be much more important for synthesis than for imputation. For synthesis, in contrast to imputation, the adequacy of the models can be checked by comparing results from the observed data to those from the synthetic data. As we illustrate below, such checks are an important feature of the *synthpop* package (Nowok et al., 2015b) for **R** (R Core Team, 2014) that we have developed for the generation of synthetic data.

1.3 Common sampling and simple vs. proper synthesis

When a gold-standard analysis is planned synthetic data must be generated by a sampling scheme that matches the one used to sample the observed data so that the same code can be used for both analyses. The simplest case is when the data can be considered as a generated by simple random sampling (SRS), as will be the case for most Census and administrative data. Synthetic data are generated by SRS and SRS methods are used for all analyses. Generation of synthetic data from data for complex surveys must match the original design. For example, if the survey was stratified then synthetic data generation should be carried out within each stratum and any sratum-specific weights will be supplied to the analyst. Methods appropriate for stratified sampling (e.g. Lumley (2004)) will be used to fit the models used to generate the syntheses, to analyse the synthetic data and for the gold-standard analysis.

We will call the condition when the synthesis matches the sample design "common-sampling". We derive alternative variance estimators for inferences from fully synthetic data that can be used when common-sampling applies. These estimators allow inferences from fully synthetic data generated from distributions with parameters fitted to the observed data without sampling the parameters from their posterior distributions. We will refer to this as "simple synthesis". It is contrary to the existing literature on synthetic data which states that fully synthetic data must be sampled from the posterior predictive distribution of the data. This is usually achieved by generating synthetic data from distributions with parameters sampled from their posterior distributions. We will refer to this as "proper synthesis". Our alternative approach (simple synthesis) to generating fully synthetic data, without the requirement to sample from the posterior distribution of the parameters, and with the alternative variance estimates described here, can be used when common-sampling applies.

1.4 Application to the UK Longitudinal Studies

The England and Wales Longitudinal Study (ONS LS) (Hattersley and Cresser (1995)), the Scottish Longitudinal Study (SLS) (Boyle et al. (2009)) and the Northern Ireland Longitudinal Study (NILS) (O'Reilly et al. (2011)) are rich micro-data sets linking samples from the national Census in each country to administrative data (births, deaths, marriages, cancer registrations and other sources) for individuals and families across several decades. Researcher access to the LSs is highly restricted due to confidentiality and legal restrictions. Thus the three LSs have a small number of users compared to other Census data products. Synthetic data with no real individuals, but which mimic the real data and preserve the relationships between variables and transitions of individuals over time, could be made available to accredited researchers to analyse on their own computers. To make such data available to users the freely available **R** package *synthpop* (Nowok et al. (2015a)) has been written as part of the SYLLS (Synthetic Data Estimation for UK Longitudinal Studies) project[1] funded by the Economic and Social Research Council to allow LS support staff to produce synthetic data for users of the LSs, which are tailored to the needs of each investigation.

1.5 Structure of this paper

The structure of this paper is as follows. The next section summarises the main theoretical results and discusses their implications for practical data analysis. Summaries of simulation studies to confirm the theoretical results are placed in the Appendix. In Section 3 we report on

[1]See http://www.lscs.ac.uk/projects/synthetic-data-estimation-for-uk-longitudinal-studies/

models and methods that have proved useful in developing the *synthpop* package as a practical tool for data synthesis and in Section 4 we present analysis of real and synthesised data from an extract taken from the SLS. The final section summarises our conclusions and points to possible future directions. Programs that can be used to reproduce the simulation studies can be obtained from the first author.

2 Methods and results

2.1 Notation and methods

Real data from a survey or a sample of census or administrative data are available to the synthesiser comprising (x_{obs}, y_{obs}) for n units where x_{obs}, which may be null, is a matrix of data that can be released unchanged to the analyst and y_{obs} is an n x p matrix of data to be synthesised. The observed data are assumed to be a sample from a population with parameters that can be estimated by the synthesiser; specifically y_{obs} is assumed to be a sample from $f(Y|x_{obs}, \theta)$ where θ is a vector of parameters. This could be a hypothetical infinite super-population or a finite population which is large enough for finite population corrections to be ignored. Note that, except in certain special circumstances, such as simulations, this assumed synthesising distribution (ASD) will not be the same as the one that can be considered as generating the observed data. The assumption that the observed data can be considered, at least for practical purposes, as a sample from the ASD is crucial to the validity of all results derived from synthetic data. We will refer to this as the ASD assumption.

The synthesiser fits the ASD, $f(Y|x_{obs}, \theta)$, to the data by a method that provides consistent estimates $\hat{\theta}$ of its parameters. For simple synthesis, data are generated as samples from this fitted distribution, $f(Y|x_{obs}, \hat{\theta})$. Proper synthesis replaces $\hat{\theta}$ with a sample from the posterior density of θ given the observed data, usually approximated by a sample $\hat{\theta}^*$ from $N(\hat{\theta}, \Omega_{\hat{\theta}})$ where $\Omega_{\hat{\theta}}$ is a consistent estimate of the variance of θ estimated from the observed data. This process is repeated m times generating m synthetic samples each of size k from the ASD and, for proper synthesis, each with a different realisation of $\hat{\theta}^*$. In most implementations of synthetic data generation, including *synthpop*, the joint distribution is defined and synthesised in terms of a series of conditional distributions as we describe in section 1.2. The generation of the synthetic data sets can then proceed in parallel to the fitting of each conditional distribution. Each column of the synthetic data is generated from the assumed distribution, conditional on x_{obs}, the fitted parameters of the conditional distribution and the synthesised values of all the previous columns of y_{obs}. A total of m synthetic data sets are generated each of which can then be considered as a sample from $f(Y|x_{obs}, \hat{\theta})$ (simple synthesis) or from $f(Y|x_{obs}, \hat{\theta}^*)$ (proper synthesis). Since a different sample from the posterior is used for every synthetic data set the totality of the proper syntheses approximates a sample from the full posterior $\int p(Y|\theta, x_{obs})p(\theta|x_{obs}, y_{obs})d\theta$ of Y, given the observed data.

The analyst may wish to make inferences for a population parameter Q using only the synthetic data. We assume that the method used for the real data will provide a consistent estimate $\hat{Q}$ of the parameter Q and a corresponding consistent estimate $U_{\hat{Q}}$ of its variance U. For simplicity results are presented here for a univariate Q, but the extension to multivariate Q follows straightforwardly, in the manner described by Reiter (2003b) and Kinney and Reiter (2010). The model of interest is fitted to each of the m synthetic data sets and yields estimates of Q as $(q_1, \cdots, q_i, \cdots, q_m)$ with estimated variances $(v_1, \cdots, v_i, \cdots, v_m)$ calculated as if each synthetic data set were real. If the ASD assumption holds then each of the q_i from a simple synthesis is a consistent estimate of $\hat{Q}$ conditional on x_{obs}. For simple synthesis this follows because any Q can be considered as a function of the parameters θ of the ASD. $\hat{Q}$ is thus the same function of $\hat{\theta}$, the parameters of the distribution which generated synthetic data. Since a function of any consistent estimator is itself consistent it follows that each q_i and their mean value $\bar{q}_m = \sum_{i=1}^{m} q_i/m$ are a consistent estimates of $\hat{Q}$ and hence of Q. For proper synthesis each q_i will be a consistent estimate of $\hat{Q}_i^*$, the equivalent fucntion of $\hat{\theta}_i^*$and hence of $\hat{Q}$ and Q since each $\hat{\theta}_i^*$ a is a consistent estimate of $\hat{\theta}_i$.

If the synthetic data are to be used to make inferences to the population then inference is required for the parameter Q and the variance of its estimate from the synthetic data. If a gold-standard analysis will be carried out, then the analyst will want to estimate the $\hat{Q}$ and its variance $V_{\hat{Q}}$ that will be obtained at the gold-standard analysis. These inferences all require the ASD assumption which can be tested by the synthesiser as part of a verification or validation step when the gold-standard analysis is carried out.

2.2 Variance estimates for fully synthetic data

The literature on synthetic data generation cited in Section 1.1 all use Raghunathan et al. (2003)'s estimator T_m for the variance of $\bar{q}_m$ as an estimate of Q from fully synthetic data produced by proper synthesis. Raghunathan et al. (2003) show that T_m is unbiased, for large

Table 1: Expected values and consistent estimates of $var(\bar{q}_m)$ for estimating of $\hat{Q}$ and Q in different settings

	Common-sampling	Type of synthesis: Simple	Type of synthesis: Proper
Estimating $\hat{Q}$			
$E[var(\bar{q}_m)]$		V/m	$(V+U)/m$
$\widehat{var(\bar{q}_m)}$	true	$T_s^{\hat{Q}} = \bar{v}_m/m$	$T_f^{\hat{Q}} = \bar{v}_m(1+k/n)/m$
$\widehat{var(\bar{q}_m)}$	false	not available	b_m/m
Estimating Q			
$E[var(\bar{q}_m)]$		$V/m+U$	$(V+U)/m+U$
$\widehat{var(\bar{q}_m)}$	true	$T_s = \bar{v}_m(1/m+k/n)$	$T_f = \bar{v}_m[(1+k/n)/m+k/n]$
$\widehat{var(\bar{q}_m)}$	false	not available	$T_m = (1+1/m)b_m - \bar{v}_m$

samples, as the common limit of two derivations; a Bayesian one with non-informative priors and a frequentist one using expectations of observed statistics. We use a simplified version of the second approach and also derive expressions for further consistent variance estimators that can be used when common-sampling applies.

The expectations of the variance of $\bar{q}_m$ as an estimate of Q for simple or proper synthesis can be written in terms of two quantities: U the expected variance of $\hat{Q}$ estimated from the observed data and V the expected variance of each q_i as an estimate of $\hat{Q}$ for simple synthesis or of $\hat{Q}^*$ for proper synthesis. In common with Raghunathan et al. (2003) we assume that the expected value of V will not vary across the q_i. For either simple or proper synthesis we can write the deviation of each q_i from Q as the sum of two independent quantities $(q_i - \hat{Q})$ and $(\hat{Q} - Q)$. The expected variance of each q_i is thus the sum of two terms the second of which is U. For simple synthesis the expectation of the first term is V. For proper synthesis we must expand the first term to two further independent terms $(q_i - \hat{Q}_i{}^*)$ and $(\hat{Q}_i{}^* - \hat{Q})$ with expected variances V and U. Averaging over the m estimates the contribution from the variance of $(q_i - \hat{Q})$ to the variance of $\bar{q}_m$ becomes V/m for simple synthesis and $(V+U)/m$ for proper synthesis because the q_i are conditionally independent given $\hat{Q}$ for both types of synthesis. These expressions are the expectations of the variance of $\bar{q}_m$ when it is used as an estimate of the $\hat{Q}$ that would be obtained from a gold standard analysis with the ASD condition met. We can think of this as the stochastic error about $\hat{Q}$ for the average results from the synthetic data. For estimating Q the term U must be added to each expected variance because $(\hat{Q} - Q)$ does not vary across syntheses.

To obtain expressions that can be calculated from the synthetic data for these expected variances we need quantities that will estimate V and U. The mean of the estimated variances of the q_i, $\bar{v}_m = \sum_{i=1}^{m} v_i/m$, can be used to estimate V for either proper or simple synthesis. Without the common-sampling condition the only estimate of U available is derived from the fact that the expected value of the between-synthesis variance of the q_i for proper synthesis, $b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1)$, is $V + U$ and this leads to the estimator T_m. But when the common-sampling condition holds U can be estimated from $\bar{v}_m$. In the simplest case where the synthetic samples are the same size as the original data, $(k = n)$, $\bar{v}_m$ estimates U as well as V, by the following argument. The variance U for estimating Q from the observed data can be written as a function of θ when the ASD condition is met. The variance of q_i estimated with v_i for a single synthetic data set will be a function of the parameters, $\hat{\theta}$ for simple synthesis, or $\hat{\theta}^*$ for proper synthesis, of the ASD that generated that synthetic data set. When the common-sampling condition is met the functions defining U and V will be identical. Since $\hat{\theta}$ and $\hat{\theta}^*$ are consistent estimates of θ it follows that each v_i and thus $\bar{v}_m$ will be a consistent estimator of U. If $k \neq n$ and variances are inversely proportional to the sample size then $\bar{v}_m k/n$ can be used to estimate U. Table 1 summarises the results and defines the new consistent variance estimates that can be used when the common-sampling conditions apply; T_s and T_f for estimating Q by simple and proper synthesis and $T_s^{\hat{Q}}$ and $T_f^{\hat{Q}}$ for estimating the gold-standard parameter $\hat{Q}$.

Drechsler (2011a) has derived an estimate for the variance of $\bar{q}_m$ when estimating Q from fully synthetic datas produced by proper synthesis, when both the real and synthetic data are generated by SRS; a special case of the common-sampling condition. This estimate becomes $T_{alt} = b_m/m + \bar{v}_m k/n$ when the finite population correction included in the original derivation is ignored. This estimator is very close to our variance estimator T_f for fully synthetic data and both avoid the problems of bias and negative values that afflict T_m. The two second terms are identical but the first terms differ. T_{alt} uses the direct estimate b_m to estimate the variance of the q_i conditional on the estimate $\hat{Q}$ from the real data, whereas T_f uses the expression $\bar{v}_m(1 + k/n)/m$. As b_m is a variance calculated from only m values we would expect T_{alt} to be less precise than T_f, although it is possible that the direct estimate in T_{alt} might be less biased in finite samples.

Note that variance estimates cannot be obtained from simple synthesis unless the common-sampling condition applies because the expectation of b_m for simple synthesis does not depend on U. But when the common-sampling condition does apply, the variance of estimates from simple synthesis will always be lower than those from proper synthesis.

All the results for Table 1 depend on the ASD assumption. However, for data generated from a sequence of conditional models, inference from synthetic data for models that are one of the conditional models will yield consistent estimates of the parameters, provided the method of estimation is consistent, even when the data were generated from a model quite different from the ASD. This follows for both proper and simple synthesis because the data consist of samples from a model with parameters that are a subset of $\hat{\theta}$ (for simple synthesis) or $\hat{\theta}^*$ (for proper synthesis). This may be a less useful property than it might first appear for two reasons. Firstly, if synthetic data are generated from the model being fitted they will contain no information to allow validity checks to assess the appropriateness of the model. Secondly, although the parameter estimates will be consistent, even when the model assumed for the real data is not correct, this will not be true of the estimates of the variance of the parameters. The variance estimates $\bar{v}_m$ are derived from the variance-covariance matrix of the predictor variables and this will not, in general, be a consistent estimate of the population variance-covariance matrix unless the ASD condition holds. All the variance estimators T_s, T_f, T_{alt}, T_m, $T_s^{\hat{Q}}$ and $T_f^{\hat{Q}}$ will be affected similarly. These properties are illustrated on the example using the SLS data in Section 4.

2.3 Results of simulations of fully synthetic data

The properties of estimates from synthetic data and their variances derived in 2.2 were confirmed by simulation studies two of which are summarised in the Appendix. The first synthesised 10,000 samples of size 100 from a multivariate Normal population and evaluated the results of fitting a linear regression model to one of the variables predicted from the others. SRS was used to generate the simulated data and to produce simple and proper syntheses from each observed data set. Two types of synthesis were evaluated, the first by generation from the joint distribution and the second by generating from conditional distributions.

The results from the first simulation showed that the estimates and their variance estimators from Table 1.1 were unbiased for this example, at least to the precision defined by the simulation. The coverage of confidence intervals was satisfactory except for the estimator T_m which took negative values for between 8 and 9% of the simulations. Results from synthesis from conditional distributions were similar except that some of the estimated coefficients from the synthetic data were slightly biased. The bias was greater for proper synthesis than for simple synthesis and was much reduced when the simulation was scaled up to samples of 1,000 rather than 100.

The second simulation used stratified sampling with appropriate estimators for the observed and synthetic data. The total sample size for each simulation was 200, in 10 strata of 20 sampled observations each. The stratum membership, which was held fixed for all syntheses can be considered as x_{obs} in this case. Again the results from Section 2.2 were confirmed. The variance estimators T_m and to a lesser extent T_f from proper sampling were slightly biased but this bias was reduced when the simulation was scaled up to samples of size 2000.

Considering that the methods we propose in our paper are intended for large sample, their performance in these small-sample simulations is very encouraging. Some estimates were slightly biased, but the bias decreased as the sample size increased, as would be expected for consistent estimators. The small biases evident in these two samples were smaller for simple synthesis than for proper synthesis.

2.4 Partially synthetic data

Since some real data are left unchanged, partially synthetic data must have the same structure as the real data and the common-sampling condition always applies usually with $k = n$, although sub-samples with $k < n$ are also possible. Reiter and Kinney (2012) present simulations which show that simple synthesis can be used for partially synthetic data with variances estimated by $T_p = b_m/m + \bar{v}_m$ as first proposed by Reiter (2003a). In terms of our development, the deviation of q_i from Q for one partial synthesis can be written as the sum of the independent terms $(q_i - \hat{Q})$ and $(\hat{Q} - Q)$, as was the case for full synthesis. The variance of the second term has expectation U which can be estimated by $\bar{v}_m$ when $k = n$. The variance of the first term can no longer be estimated by $\bar{v}_m$ because the synthetic data are no longer generated from $f(Y|x_{obs}, \hat{\theta})$ because only part of the data is replaced. Thus this first term must use the direct estimate b_m of the variation between the syntheses. This term will be a minor contribution to T_p if only a small fraction of the data are replaced so the fact that it may be based on a small number of degrees of freedom may be less important than for full synthesis.

3 Practical aspects of data synthesis

Real survey, census or administrative data may bear very little resemblance to the models used to derive the theory of synthetic data. Continuous data may have distributions that are nothing like a Normal distribution, even after a suitable transformation. Categorical data may have many complex interactions that it would be unreasonable to expect the synthesiser to investigate. Furthermore, real data may be subject to constraints that must be respected for the survey data. For example, if an analyst were presented with synthetic data on children in families where a natural parent was less than 12 years older than a child, the utility of the data would be questionable.

Fortunately these questions have been addressed for synthetic data and the literature contains many options, some of which have been implemented in the *synthpop* package. Woodcock and Benedetto (2009) describe and evaluate methods that preserve the marginal distributions of continuous variables and these can also be adapted to include an element of smoothing to prevent the identification of unique values. A number of methods from machine learning have been used successfully to generate synthetic data (Reiter (2005b), Caiola and Reiter (2010), Drechsler and Reiter (2011)). These methods are adaptive and may be able to reproduce the main features of the data without the need for exploratory analysis. Classification and regression tree models (CART), which performed well in the evaluation carried out by Drechsler and Reiter (2011), can be selected to synthesise data with *synthpop* and they are the default method if no detailed models are specified. The use of a sequence of conditional distributions makes it easy to incorporate constraints on data values for synthetic samples. Variables that define the constraints must be synthesised first and the constrained variable is then synthesised with the constraint satisfied.

When data to be synthesised have missing values they could be replaced by imputation and the mutiply imputed data sets can then be synthesised. This approach has been illustrated by Drechsler (2011b) and US Census Bureau (2013) who use formulae and variance estimates for combining multiple imputations with multiple syntheses due to Reiter (2004). We have not adopted this approach in *synthpop* because we expect that the choices about handling missing data for a particular project should be the responsibility of the analyst. Using a missing-at-random approach we synthesise the missingness indicator first, and then synthesise the remaining cases from a fit to the non-missing cases in the observed data. Both the synthesised values and the missingness indicator can then be used together in the synthesis of later variables. This guarantees that any relationships with the missingness indicator are maintained in the synthetic data. For variables earlier in the sequence this is assured by the model that predicts missingness and, for those later in the sequence, by having the missingness indicator as one of their predictors. An analyst can use synthesised data with missing values to decide how to handle them and their methods can readily be run on the real data. In some cases further synthetic data, with missing values ignored or imputed, could be provided to the analyst.

Data from the LSs often includes time to event data. These may be defined as a series of dates or as a follow-up time and an indicator of the event at the end of follow up. For example, the LSs are linked to death registrations and emigration records. To synthesise such data the event indicator is synthesised first and the follow-up times are synthesised separately for each type of event. Possible models for follow-up times are parametric survival models (Weibull or log-Normal) or a CART method applied to Kaplan-Meier survival estimates (Hothorn et al. (2006)). Cox proportional hazard models would be more difficult to fit because they would require the vector of all observed event times and the corresponding baseline hazard to be used in simulating the synthetic data. Poisson models can be used for person-years analyses.

4 Example

4.1 Methods

The fitting of formal models is only a small part of any statistical analysis. The majority of an analyst's time is taking up with checking and exploring the data and in carrying out preliminary tabulations. We aim to produce synthetic data that can be used for this type of analysis. To illustrate synthesis of LS data we have extracted data on age, sex, marital status, ethnic group and long-term illness from the SLS database for the 1991 and 2001 Censuses. The acronyms AGE9, SEX9, MSTAT9, ETH9, ILL9 and AGE0, SEX0, MSTAT0, ETH0, ILL0 are used to describe them. The synthesis was carried out for over 186 thousand SLS members who were present at both Censuses. No preliminary data cleaning was carried out on the extract. Some variables had a small percentage of missing values. For the categorical variables the missing cases are simply handled as an additional category, but AGE0 had 0.15% of missing cases coded as -999 and this formed a missingness indicator for AGE0. Exploratory analyses of univariate distributions, cross tabulations and results from fitted models were compared for

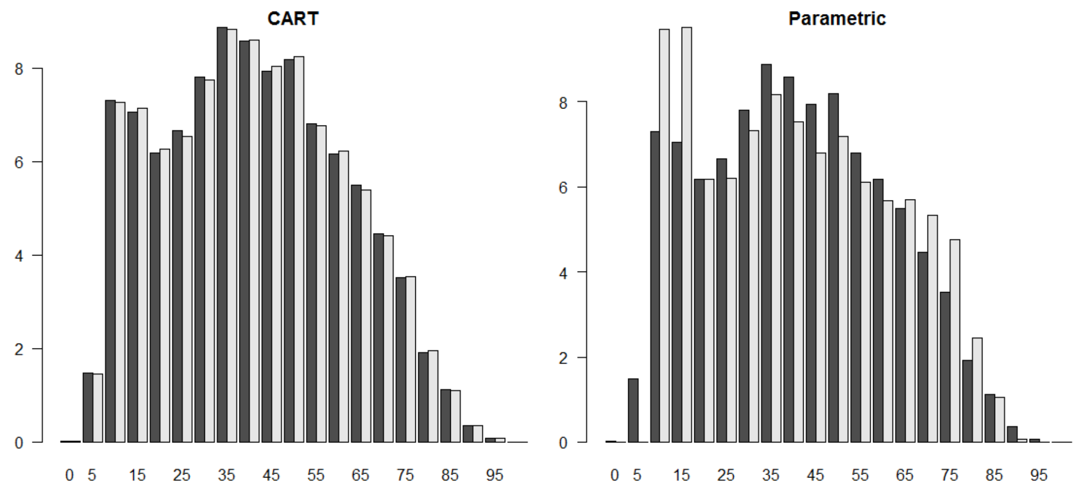

Figure 1: Comparison of real (black) and synthetic (grey) data for AGE0 in 5 year age groups

the real and the synthetic data. The *synthpop* package was used for all analyses and more details of the methods can be found in Nowok et al. (2015a).

Initial analysis used simple syntheses with two choices of models. The first (CART) used CART for all the variables and the second (Parametric) used an appropriate parametric method for each variable: polychotomous or logistic regression for categorical data and distribution-preserving linear regression for AGE0[2] A lower triangular prediction matrix was used in both cases so that all previously synthesised variables were used in the predictions and $m = 50$ syntheses were produced for each. The ordering of the variables used in the final syntheses and for the results presented here was ETH9, ETH0,AGE9, SEX9, MSTAT9, ILL9, AGE0, SEX0, MSTAT0, ILL0.

4.2 Results of exploratory analyses

A problem with the initial run of the synthetic data was that some SLS members under 16 had marital status"married" in the synthetic data, with the number of cases being larger for the Parametric syntheses. This was readily fixed by imposing this rule during the syntheses. Marginal distributions of all the variables were comparable to those for the real daa for syntheses produced from CART, but for parametric methods, synthetic data for two variables

[2]The data were transformed to Normal scores calculated from their percentiles and regression carried out on these values. The predicted values were then mapped back to the original data with the reverse transformation. A smoothing option can be used for this method, but was not necessary here.

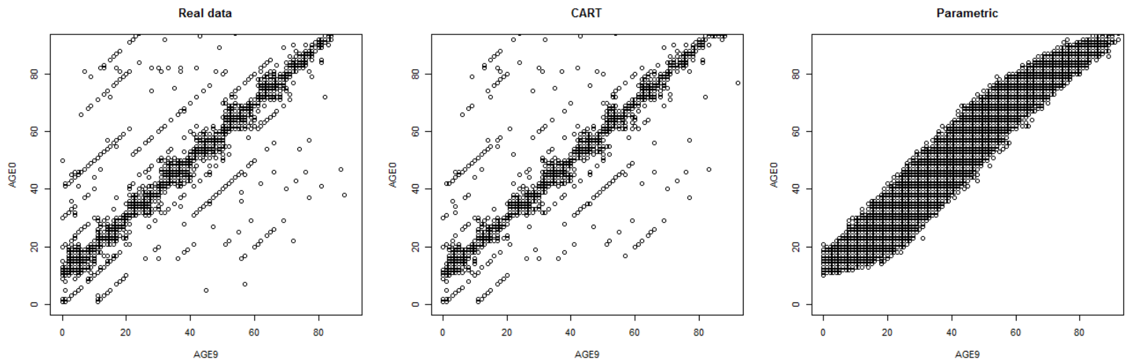

Figure 2: Plots of AGE9 against AGE0 for real data and one synthesis by each of CART and parametric methods (most points refer to multiple observations)

had marginal distributions that did not match the observed data. Initially the variables ETH9 and ETH0 were placed later in the synthesis order. This produced unsatisfactory results for small ethnic groups, when parametric methods were used. The most extreme differences were for Bangladeshis, the smallest category with only 28 SLS members in 2001, who were over-represented with an average of over 4,000 in the synthetic data sets. Exploratory analyses showed that these problems arose from attempting to fit a polychotomous regression to sparse data with many zero cells, when some parameters are fitted at their extremity. Moving the two ethnic group variables to the start of the Parametric synthesis overcame the problem. Another approach of defining a predictor matrix with smaller models used to predict the ethnic group variables also gave satisfactory results.

The other variable that was affected was AGE0 (see Figure 1). From Figure 2 we can see why the parametric method failed to reproduce the distribution of AGE0[3]. Age is recorded in full years and, in most cases, AGE0 is exactly AGE9+10. Exceptions could be those with birthdays between the dates of the two Censuses as well as various data errors or mismatches. Differences of exactly 10 or 20 years are common. The parametric method could not reproduce this pattern. The parametric syntheses were rerun with the method for just this variable changed to CART giving satisfactory results.

Thus for these exploratory analyses the CART method gives more satisfactory results than parametric methods, and with no requirement to customise the analyses in any way.

[3]Points at ages over 90 are not shown to avoid any possible disclosure of extreme ages.

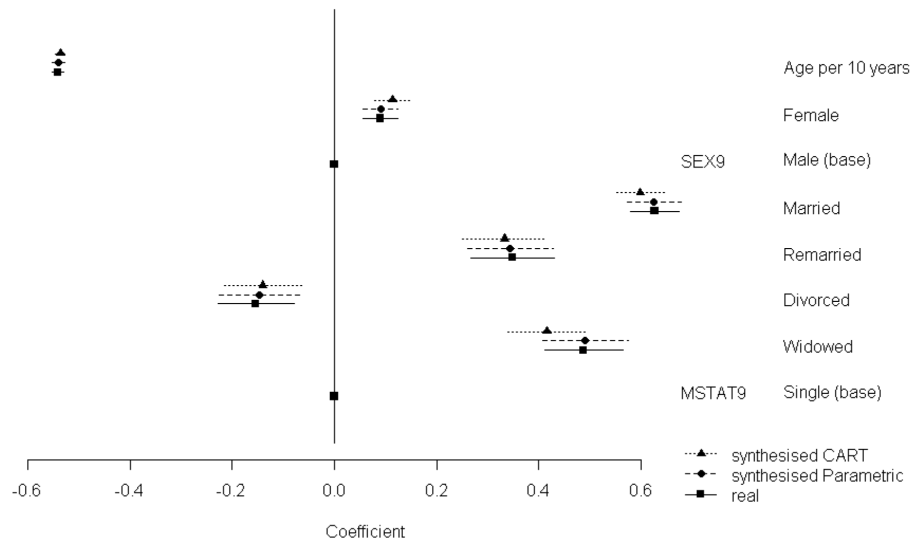

Figure 3: Coefficients of fit to ILL9="No" from AGE9, SEX9 and MSTAT9 for real and synthetic data

4.3 Results of fitting models to data generated by simple synthesis

The difference between CART methods and parametric methods might be expected to be different for fitted models. This was investigated by fitting a logistic regression to absence of long-term illness in 1991 (ILL9) as a linear model from AGE9, MSTAT9 and SEX9. A smaller synthesises were carried out with $m = 10$ and just these four variables with ILL9 as the last variable synthesised so as to investigate a model that was part of the parametric syntheses.

Initially simple synthesis was used by parametric and CART methods with the rule for MSTAT9 for the under 16s in place. Figure 3 compares the estimates from the real data with the averages from the 50 simple syntheses from parametric and CART models. Confidence intervals here are calculated from $\bar{v}_m$ the estimate of the variance that would be expected from the observed data. As expected, the estimates from the parametric models are very close to the real estimates, and a formal test showed that there was no evidence of any bias In this case the estimates from the CART syntheses are also fairly close and certainly would be satisfactory for exploratory analyses. We can see that freedom from long-term illness decreases sharply with age and is higher for females than males. Adjusting for age and sex, those married, remarried or widowed are more likely to be free from long-term illness than those who are single, whereas

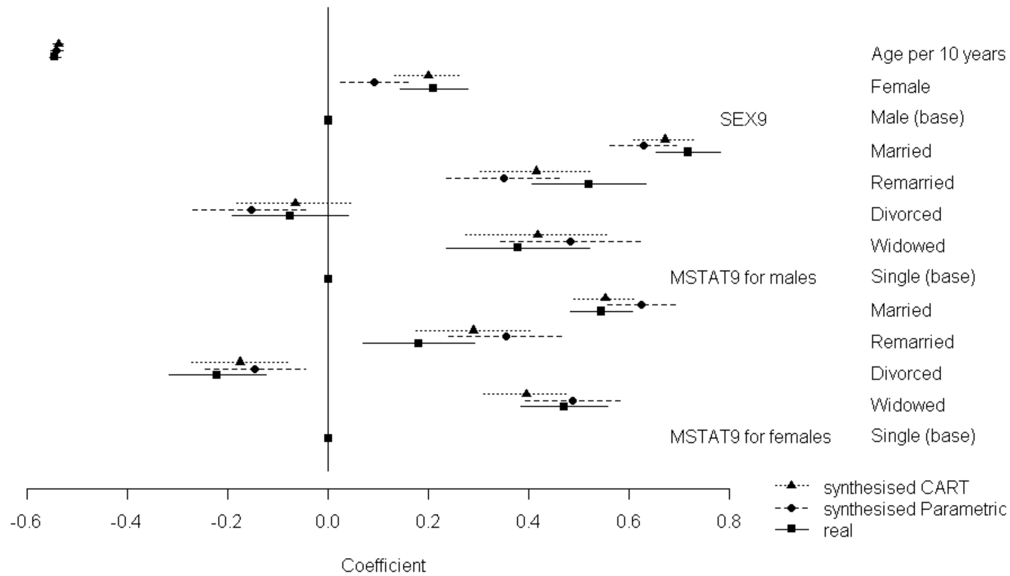

Figure 4: Coefficients of fit to ILL9="No" from AGE9, SEX9 and MSTAT9*SEX9 interaction for real and synthetic data

the opposite is true for the divorced.

The standard errors calculated from $\bar{v}_m$ were compared to those obtained from the observed data. All the estimates from the synthetic data were close to those from the observed, the largest ratio between the two being only 1.12. Rather surprisingly, we found that estimates of standard errors from inferences based on CART syntheses had almost no evidence of bias, whereas there was evidence of a small bias for those from parametric syntheses.

A further disadvantage of parametric methods is that they preclude an analyst, with access only to the synthetic data, from checking departures from an assumed model, such as lack of linearity or the absence of interactions. This is illustrated here by fitting a further model which includes a sex by marital status interaction. Results are shown in Figure 4. For the real data there is evidence of an interaction. The association of being married with lack of illness is stronger for men than for women. The CART syntheses do a reasonable job of reproducing this, whereas the parametric syntheses show no evidence of this interaction since they are generated from an interaction-free model.

These results all assume that the analyst is interested in estimating the results which might be obtained from the real data, rather than in making inferences to population parameters directly from the synthetic data. We believe this is an appropriate use of synthetic data because we can never be completely assured of the validity of inference from such data. Further experience with synthetic data may show that this is too cautious a position, but it is a safe one for now. In the next section we present inferences for the population parameters by simple and proper synthesis for this example.

4.4 Inference to population parameters from data generated by simple and proper synthesis

Finally, we fitted the same model to data generated by proper synthesis with parametric and CART methods and compared variance estimates with those from simple synthesis. A proper version of CART synthesis starts by taking a bootstrap sample of the real data and fitting the CART models to this sample. For simple synthesis we have a single variance estimate T_s whereas, for proper synthesis, we have three variance estimates T_f, T_{alt} and T_m. Proper syntheses with parametric and CART models were carried out with $m = 10$ in each case.

The synthetic estimates of the coefficients from proper synthesis showed the same patterns for CART and parametric methods as in Figure 3. The expected values of the variances of $\bar{q}_m$ as an estimate of Q should be $V(1 + 1/m)$ for simple synthesis and $V(1 + 2/m)$ here, since $k = n$. Table 2 gives the ratios of the standard errors from simple and proper synthesis divided by $\sqrt{V_{\hat{Q}}}$, the standard error from the observed data, as an estimate of $\sqrt{V}$. If the ASD assumption holds the expected value of these ratios of standard errors would be approximately 1.049 for simple synthesis and 1.095 for proper synthesis. For the CART syntheses, which gave approximately unbiased standard errors for estimating $\hat{Q}$ we can see that standard errors calculated from T_s and T_f are close to their expectations. Those calculated from T_{alt} are somewhat larger for the CART methods. The results are more irregular for the parametric methods which gave more biased estimates of the standard errors for simple synthesis. In both cases, as expected, the estimator from T_m is quite unsatisfactory, giving a negative value in one case.

These results support our recommendation of the use of the variance estimates T_s and T_f for estimating Q from data generated by simple and proper synthesis, respectively.

5 Summary and future directions

We have shown that simple synthesis can provide valid inferences for fully synthetic data in the situation when the common-sampling condition is satisfied. When this is the case simple

Table 2: Standard errors for estimating the population coefficients of a logistic regression of ILLP9 predicted from AGE9, SEX9 and MSTATT9 from simple and proper synthetic data produced by parametric and CART methods. Standard errors are calculated from different variance estimates and expressed as proportions of the standard errors from the observed data

	Ratios to standard error from the real data							
	Parametric				*CART*			
	Simple synthesis	*Proper synthesis*			*Simple synthesis*	*Proper synthesis*		
	$\sqrt{T_s}$	$\sqrt{T_f}$	$\sqrt{T_{alt}}$	$\sqrt{T_m}$	$\sqrt{T_s}$	$\sqrt{T_f}$	$\sqrt{T_{alt}}$	$\sqrt{T_m}$
Intercept	1.046	1.091	1.078	0.940	1.045	1.098	1.107	1.201
AGE9	1.165	1.217	1.196	0.952	1.048	1.095	1.133	1.458
SEX9								
Female	1.043	1.089	1.174	1.820	1.049	1.095	1.131	1.447
MSTATT9								
Married	1.172	1.225	1.248	1.465	1.048	1.097	1.232	2.159
Remarried	1.086	1.135	1.103	0.704	1.052	1.095	1.119	1.337
Divorced	1.091	1.143	1.081	NA	1.064	1.089	1.212	2.076
Widowed	1.145	1.199	1.208	1.294	1.048	1.099	1.216	2.044

NA Negative variance estimate

synthesis will be preferred as estimates from simple synthesis will always be closer to the real data than those from proper synthesis. We have also derived variance estimates that can be used when this condition holds that have better properties than the most commonly used variance estimates for fully synthetic data. These results are valid for large samples with methods of estimation that provide consistent estimates of parameters and their variances.

The results in Section 4 recommend non-parametric methods for this example from the SLS. The Parametric methods were less satisfactory at reproducing the marginal distributions of some variables. More importantly, although the parametric model gave a better fit than CART to a model that was part of the syntheses, CART was better at detecting an interaction that was not modelled by the parametric fit that generated the synthetic data. So parametric models may be better at getting precise results for certain models, but CART's less precise fits may be more robust for picking out features of the data that had not been anticipated by the synthesiser.

More experience is needed on the best way to carry out syntheses. Recommendations are needed on choosing the ordering of variables during synthesis, for deciding whether reduced models excluding some variables should be used and for fine-tuning the parameters of CART models. The *synthpop* package is intended to facilitate this and new methods can be added by the user that are not at present part of the package.

The synthesiser, with access to the real data, ought to carry out checks on the validity of the data before it is released to the analyst. At a minimum, a visual check on all the marginal distributions should be carried out. The *synthpop* package includes a function to do this and Figure 3 is an example of part of its output. We also hope that experience from users of the LSs who compare their results from the real data, after preliminary analyses of the synthetic data, will help to develop best practice. This should include code to verify any assumptions (such as the absence of interactions) in the fitted models since these might not be evident with the synthetic data.

We are not addressing questions of disclosure control in detail in this paper, but a few comments are relevant. For fully synthetic data no records refer to real cases, so disclosure of a real person is unlikely, but not zero, so further evaluation of the disclosure risk from simple synthesis needs to be undertaken. Partially synthetic data raise more problems as individuals may be identifiable and their synthesised data may be inferred from an analysis of the synthesised values across syntheses. Reiter and Kinney (2012) argue that this risk is lower for simple synthesis of partially synthetic data because fewer synthetic data sets are required. Another aspect of disclosure control is of more concern to data providers. If an intruder sees the synthetic data they may mistakenly believe it to be real and attempt an identification, with subsequent loss of reputation for the data collection agency. We are addressing these concerns by restricting access to synthetic data to trained and accredited researchers and also by adding labels to the synthetic data sets to make it clear that the data are fake.

6 Acknowledgements

We are grateful to the staff of the Longitudinal Studies Centre – Scotland (LSCS) for their help in providing the data analysed here and in supporting our work in developing the *synthpop* package. The LSCS is supported by the ESRC/JISC, the Scottish Funding Council, the Chief Scientists Office and the Scottish Government. The authors alone are responsible for the interpretation of the data. Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queens Printer for Scotland. We would like to thank the editors and referees who have helped us to improve the presentation of our work.

References

Arnold, B., E. Castillo, and J. M. Sarabia (2001). Conditionally Specified Distributions: an Introduction. *Statist. Science 16*, 249–265.

Boyle, P., P. Feijten, Z. Feng, L. Hattersley, Z. Huang, J. Nolan, and G. Raab (2009). Cohort Profile: the Scottish Longitudinal Study (SLS). *Int. J. Epidem. 38*, 385–392.

Caiola, G. and J. P. Reiter (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy 3*, 27–42.

Drechsler, J. (2011a). Improved variance estimation for fully synthetic datasets. Joint UNECE/Eurostat work session on statistical data confidentiality 2011. Working paper, 18. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/18_Drechsler.pdf. Accessed: 2014-07-14.

Drechsler, J. (2011b). *Synthetic Data Sets for Statistical Disclosure Control.* Springer, New York.

Drechsler, J. and J. P. Reiter (2011). An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets, computational statistics and data analysis. *Computational Statistics and Data Analysiss 55*, 3232 – 3243.

Dreschler, J. (2012). New data dissemination approaches in old Europe - synthetic datasets for a German establishment survey. *J. Appl. Stats. 39*, 243–265.

Dreschler, J. and J. P. Reiter (2010). Sampling with synthesis: a new approach for releasing public use census midrodata. *J. Amer. Statist. Ass. 105*, 1347–1357.

Elliot, M. and K. Purdam (2007). A case study of the impact of statistical disclosure control on data quality in the individual uk samples of anonymized records. *Env. Plan. A 39*, 1101–1118.

Hattersley, L. and R. Cresser (1995). *The Longitudinal Study, 1971-1991: History, organisation and quality of data. LS Series no.7.* London: The Stationery Office.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: a onditional inference framework. *Journal of Computational and Graphical Statistics 15*, 651–674.

Kinney, S. K. and J. P. Reiter (2010). Tests of multivariate hypotheses when using multiple imputation for missing data and disclosure limitation. *J. Off. Stats. 26*, 301–315.

Lumley, T. (2004). Analysis of complex survey samples. *J. Statist. Software 9*, 1–19.

Nowok, B., G. M. Raab, and C. Dibben (2015a). synthpop : Bespoke creation of synthetic data in R. Package vignette `http://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf`. Accessed: 2015-02-26.

Nowok, B., G. M. Raab, and C. Dibben (2015b). *synthpop: Generating synthetic versions of sensitive microdata for statistical disclosure control*. R package version 1.1-0.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review 57*, 1701–1775.

O'Reilly, D., M. Rosato, G. Catney, F. Johnston, and M. Brolly (2011). Cohort description: The Northern Ireland Longitudinal Study (NILS). *Int. J. Epidem. 41*, 634–641.

R Core Team (2014). *"R: A Language and Environment for Statistical Computing"*. Vienna, Austria: R Foundation for Statistical Computing.

Raghunathan, T., J. Lepkowski, J. van Hoewyk, and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodol. 27*.

Raghunathan, T. E., J. P. Reiter, and D. B. Rubin (2003). Multiple imputation for statistical disclosure limitation. *J. Off. Stats. 19*, 1–17.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *J. Off. Stats. 18*, 531–544.

Reiter, J. P. (2003a). Inference for partially synthetic, public use microdata sets. *Survey Methodol. 21*, 181–188.

Reiter, J. P. (2003b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *J. Statist. Plan. Inf. 131*, 365–77.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodol. 30*, 235–242.

Reiter, J. P. (2005a). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *J. R. Statist. Soc. A 168*, 185–205.

Reiter, J. P. (2005b). Using CART to generate partially synthetic, public use microdata. *J. Off. Stats. 21*, 441–462.

Reiter, J. P. and S. K. Kinney (2012). Inferentially valid, partially synthetic data: generating from posterior predictive distributions not necessary. *J. Off. Stats. 28*, 583–590.

Reiter, T. J. P. and E. Raghunathan (2007). The multiple adaptations of multiple imputation. *J. Amer. Statist. Soc. 102*, 1462–1471.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley,New York.

US Census Bureau (2011). Synthetic Longitudinal Business Database. Washington DC; Cornell University, Synthetic Data Server [distributor], Ithaca, NY, 2011. `http://www.census.gov/ces/dataproducts/synlbd/methodology.html`. Accessed: 2015-02-25.

US Census Bureau (2013). SIPP Synthetic Beta: Version 5.1 [Computer file]. Washington DC; Cornell University, Synthetic Data Server [distributor], Ithaca, NY, 2013. `http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.h` Accessed: 2015-02-25.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statist. Meth. Med. Res. 16*, 219–242.

Winkler, W. E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytical properties are not justified. Technical Report Series 2007 21, Statistical Research Division, U.S. Census Bureau, Washington, DC 20233.

Woodcock, S. D. and G. Benedetto (2009). Distribution-preserving statistical disclosure limitation. *Comp. Stats. Data Anal. 53*, 4228–4242.

A Appendix

The code used to run these simulations in **R** is available as supplementary material. Four files are available, The first includes functions required for the simulations which must be run before the others. Two files carry out the simulations in A.1, one for simulation from the joint distribution, one for simulation from the conditional distributions and there is a single file for the simulations in A.2.

A.1 Simulation to evaluate methods for synthetic data with simple random sampling

This simulation is based on one used to evaluate methods for synthetic data and the variance estimator T_m by Raghunathan et al. (2003). The population was created by drawing a sample of size $N = 10,000$ from $N(0, \Sigma)$, where Σ is a 5 by 5 matrix with diagonal elements 1 and off-diagonals 0.5. The columns of the population are denoted by $(y_1, y_2, y_3, y_4, y_5)$. Each of 10,000 simulations selected a random sample of size $n = 100$ from this population as the observed data. The parameters of the five-variate Normal distribution were estimated for each and two sets of $M = 5$ synthetic data sets of size $k = 250$ were generated, one by simple synthesis and one by proper synthesis. This was identical to the original simulation except that the number of simulations was increased from 500 to 10,000 and the population size was increased from 1,000 to 10,000. With a population of 1,000 the finite population correction factor is 0.9, but the syntheses, as described, were not generated from finite populations.

A linear model for y_1 predicted from the other 4 variables was fitted to each simulated data and to the two sets of synthetic data.. For population inference the parameters Q are the coefficients of this model fitted to the population. For inference to the result that would be obtained from the observed data the targets are $\hat{Q}$, the parameters obtained by fitting the model to each simulated observed data set of size $n = 100$. In all cases the estimate from the synthetic data is the mean of the estimates for each of $M = 5$ synthetic samples. The estimates presented in the lower half of Table 1 were computed for each set of syntheses and the coverage of nominal 95% intervals computed for each.

A second simulation was carried out to evaluate synthesis from fits to conditional distributions. Simulated data were generated as above, but the synthesis was carried out from fits to the conditional distributions in the order y_1, y_2, y_3, y_4, y_5. The distribution used to generate the simulations has 15 parameters, while the conditional model has 20, and fits by these two methods will not be identical for finite samples. However, these 20 parameters are all functions of the original 15. Thus our large sample results hold and two approaches should give the same results for large samples'. The results for the fitted model were evaluated in the same way as for the joint distribution.

A.1.1 Results for syntheses from the joint distribution

The synthetic data gave unbiased estimates of the coefficients of the linear model and the estimators T_s for simple synthesis and T_f and T_m for proper synthesis gave unbiased estimates of their variances. However, T_m was negative in for between 8 and 9% of cases. Two approaches were used to adjust the coverage for the negative values of T_m; an estimator ($T_m adj$), proposed by Reiter (2002) or basing the intervals on only the cases where ($T_m > 0$). Table 3 shows the estimated coverage calculated for all these variance estimates. We can see that all the results are very satisfactory except those based on T_m, which have reduced coverage. The variances using T_m was less precise than the others with a variance between 20 and 50 times greater than that for T_f.

Table 3: Coverage, calculated from the simulation from the joint distribution, of 95 % confidence intervals for Q or $\hat{Q}$ calculated from different variance estimators

		Estimating Q			Estimating $\hat{Q}$	
	T_s	T_f	$T_m adj$	$T_m > 0$	$T_s^{\hat{Q}}$	$T_f^{\hat{Q}}$
intercept	94.50	94.70	85.70	85.10	95.00	94.60
y2	94.80	94.90	86.50	85.90	95.00	94.80
y3	94.50	94.70	86.40	85.80	94.90	94.90
y4	94.30	94.50	86.00	85.60	94.60	94.50
y5	95.00	94.80	86.20	85.60	95.20	94.60

A.1.2 Results for syntheses from the conditional distributions

Coverage results were very similar to those from the joint distribution in Table 3, but there was evidence of a small bias in the estimates, which was most pronounced for the coefficients of y_4 and y_5 in the fitted model (Table 4 with n = 100). The bias was absent when the order of the sequence in which the variables were synthesised was reversed as would be expected because the fitted model is now the same as one of the conditional models used in the synthesis. Note that the bias is greater for proper than for simple synthesis. However, these biases are smaller than the standard errors of the estimates. The largest bias for estimating Q is 0.12 times its standard error and for estimating $\hat{Q}$ 0.3 times. This explains the coverage results since such small shifts in a mean will have only a small effect on the coverage. When the simulation was scaled up by a factor of 10 the biases were much reduced (Table 4 with $n = 1,000$).

Table 4: Results for synthesising via conditional distributions. Mean bias of estimates of Q and $\hat{Q}$ from 10,000 simulated samples and estimates calculated from the corresponding synthetic data

	Estimating Q			Estimating $\hat{Q}$		
	Population	bias x 1000		Mean of	bias x 1000	
	values	simple	proper	sample values	simple	proper
n = 100						
intercept	-0.005	0.137	0.078	-0.005	0.414	0.355
y2	0.284	0.051	-1.130	0.285	-0.911	-2.092
y3	0.149	-1.778	-3.507	0.149	-2.090	-3.819
y4	0.202	-8.807	-14.835	0.200	-6.611	-12.639
y5	0.179	-5.894	-13.026	0.181	-7.792	-14.925
n = 1,000						
ine intercept	0.001	-0.071	-0.240	0.001	-0.016	-0.186
y2	0.200	0.269	0.496	0.200	0.292	0.519
y3	0.200	-0.365	-0.444	0.200	-0.127	-0.206
y4	0.198	0.481	-0.058	0.199	-0.246	-0.785
y5	0.204	-1.057	-1.599	0.203	-0.757	-1.299

A.2 Simulation to evaluate synthesis for a stratified sample

A.2.1 Methods

This simulation was based on that presented in section 3.2 of Reiter (2002) where the simulated observed data and the synthetic data are drawn as stratified samples. A population of size N = 10,000 was created consisting of 10 strata $h = 1, 2, ..., 10$ each of size $N_h = 1,000$ and where the variable Y of interest is distributed as $N(10h, h^2)$, in the h^{th} stratum. Each simulation selected a stratified random sample of total size $n = 200$, with $n_h = 20$ from each stratum, from this population to represent the observed data. Synthetic samples of size 200 were then generated. In Reiter (2002) this was done by first taking a random sample from the population without constraining it to balance the population totals. In the simulation reported here exactly $n_h = 20$ values were synthesised for each stratum so that the common-sampling condition holds. The information on stratum membership, which is fixed for each synthetic sample is the x_obs for this example.

Two sets of $m = 100$ syntheses were generated by proper synthesis and by simple synthesis. The parameter (Q) to be estimated is the population mean estimated by the stratified estimate $\sum_{h=1}^{10} N_h/N\bar{y}_h$ where $\bar{y}_h$ is the mean for of the observations $y_{hj}, j = 1, 2, .., 10$ in the h^{th} stratum with variance estimated from (1).

$$\sum_{h=1}^{10} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2}{(n_h - 1)\, n_h} \tag{1}$$

The stratified sample estimate ($\hat{Q}$) of the mean was calculated for each simulation. For each set of syntheses the average and variance of the stratified sampling estimates of the mean and its corresponding variance estimate from (1) were calculated to give $\bar{q}_m$, b_m and $\bar{v}_m$. The variance estimates for synthetic data, T_s, $T_s^{\hat{Q}}$ for simple synthesis and T_f, $T_f^{\hat{Q}}$ and T_m for proper synthesis, were calculated as functions of these. Confidence intervals coverages were calculated based on each of the variance estimates. This sample design has an extreme design effect. If the usual SRS formula for the variance of the mean were to be used the result would be more that 22 times greater than if the correct one were used.

A.2.2 Results

Table 5: Comparison of variances of synthetic estimates of the mean from the simulations for stratified sampling with the mean values of their estimates and the corresponding coverage of 95% confidence intervals

	Variance from the simulation	Mean of estimates of variance	Coverage of 95% interval
Observed data	0.1899		
Estimating Q			
Proper synthesis	0.1962	$T_f = 0.2165$	95.62
		$T_m = 0.2259$	95.17
Simple synthesis	0.1933	$T_s = 0.1918$	94.64
Estimating $\hat{Q}$			
Proper synthesis	0.0044	$T_f^{\hat{Q}} = 0.0042$	94.57
Simple synthesis	0.0019	$T_s^{\hat{Q}} = 0.0019$	94.64

Both simple and proper synthesis gave unbiased estimates of Q and $\hat{Q}$. Because $m = 100$

syntheses were obtained for each simulated observed sample the estimator T_m was negative in only 1 out of 10,000 simulations, and this value was simply ignored.

Table 5 compares the variances from the simulations with the mean of the estimators from Table 1 as well as the coverage estimates of nominal 95% intervals for each. The variance estimate T_m from proper synthesis for Q is slightly biased upwards, as was found in the original publication. The estimate T_f for proper synthesis was also slightly biased, but to a lesser extent than T_m, whereas T_s for simple synthesis appears unbiased. This bias was much reduced when the whole simulation was scaled up by a factor of 10 to give a sample of 200 in each stratum, rather than 20 as in the original publication. The coverage of the confidence intervals was satisfactory for all the estimators.